Usage des bandes hors Tier 1

L. Duflot

Introduction

- Un système de bande est un stockage
  - Possiblement moins coûteux que le disque (dépend des performances attendues)
  - Probablement moins consommateur d'énergie
- Le but est de faire une mini revue de l'utilisation "non standard" de bandes, c'est à dire hors du modèle des bandes au Tier 1 comme stockage de grande volumétrie avec stagein/out organisé et à grande bande passante
- Clairement pas d'application immédiate dans LCG-France pour les sites Tier 2 mais à voir comme une exploration "think outside the box"
- Je connais deux T2 ayant des bandes : NET2 (ATLAS – Saul Youssef) et MIT (CMS – même Data Center NESE ?)
- BNL a fait une expérimentation de « service HSM » MÁS (BNLLAKE) (Éric Lançon)
- Pour les aspects opérationnels je serai très biaisé ATLAS vu les sites que j'ai contacté et mon expérience perso

Bandes à NESE

MGHPCC

Boston University
Harvard University
MIT
Northeastern University
University of Massachusetts

- 15 megawatt $90M single purpose data center
- Near zero Carbon footprint
- Space power and cooling for 780 racks
- More than 300,000 x86 cores, millions of gpu cores
- 100Gb/s multi-fiber ring to Boston, NYC and Albany
- Three new top500 in the past year
- Located in Holyoke, MA
- Thousands of researchers in all fields
- 200,000 students; 35% of degrees awarded in MA

NESE Ceph

- 32 PB Ceph
- 80% buy-in
- Rapidly expanding: 167 projects, 162 PIs, 112 organizations
- 9.5 PB raw NESE_DATADISK, CephFS, 8+3 EC

NESE Tape(*New*)

- 157 PB capacity IBM TS4500 tape library, 1 PB GPFS cache
- 110.4 PB buy-in already
- 50 PB for ATLAS
- 10 PB for CMS to start…

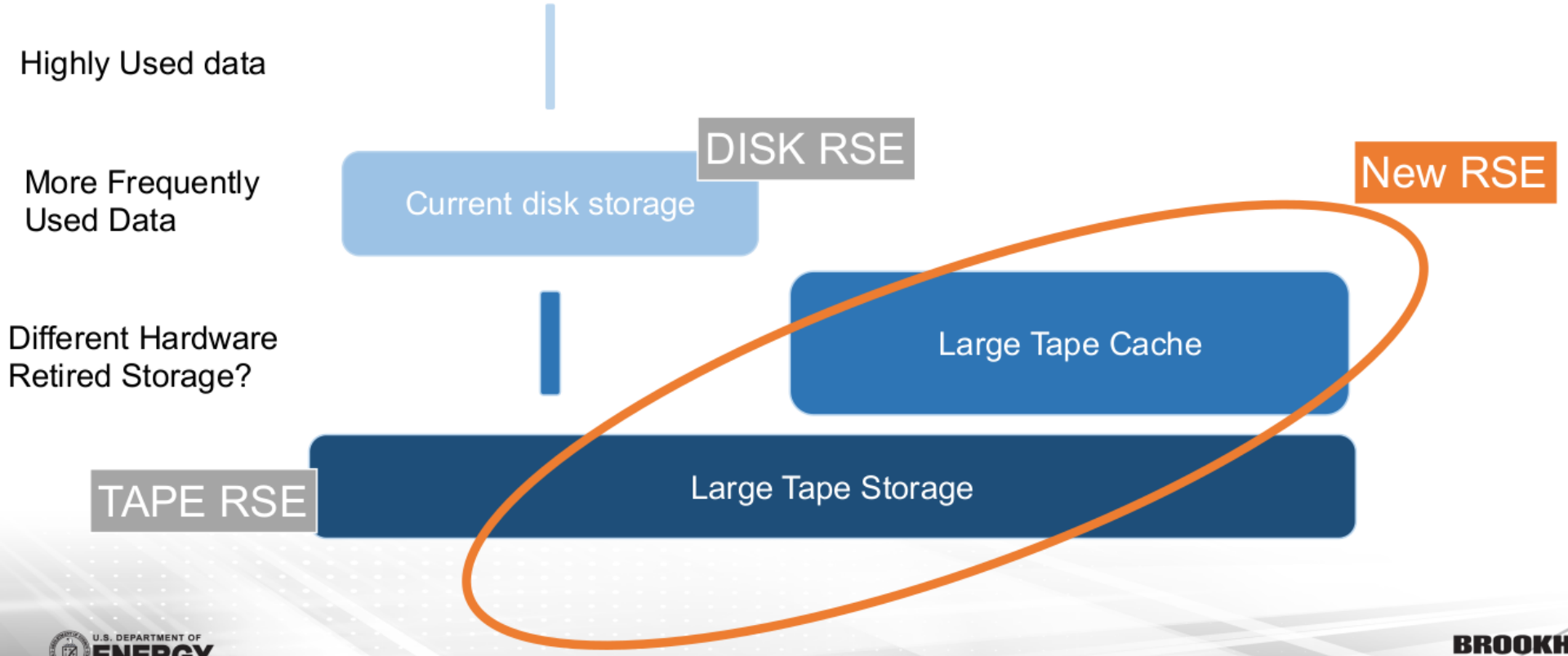
Saul Youssef

- Prévoit d'utiliser l'interface xrootd au bande comme pour CTA au CERN
- Première allocation de 10PB pour ATLAS pour archivage de DAOD
- Pas d'autre type d'utilisation prévue pour le moment

MÁS at BNL

BNLLAKE

MÁS implementation



U.S. DEPARTMENT OF ENERGY

15

BROOKHAVEN NATIONAL LABORATORY

Éric Lançon

MÁS

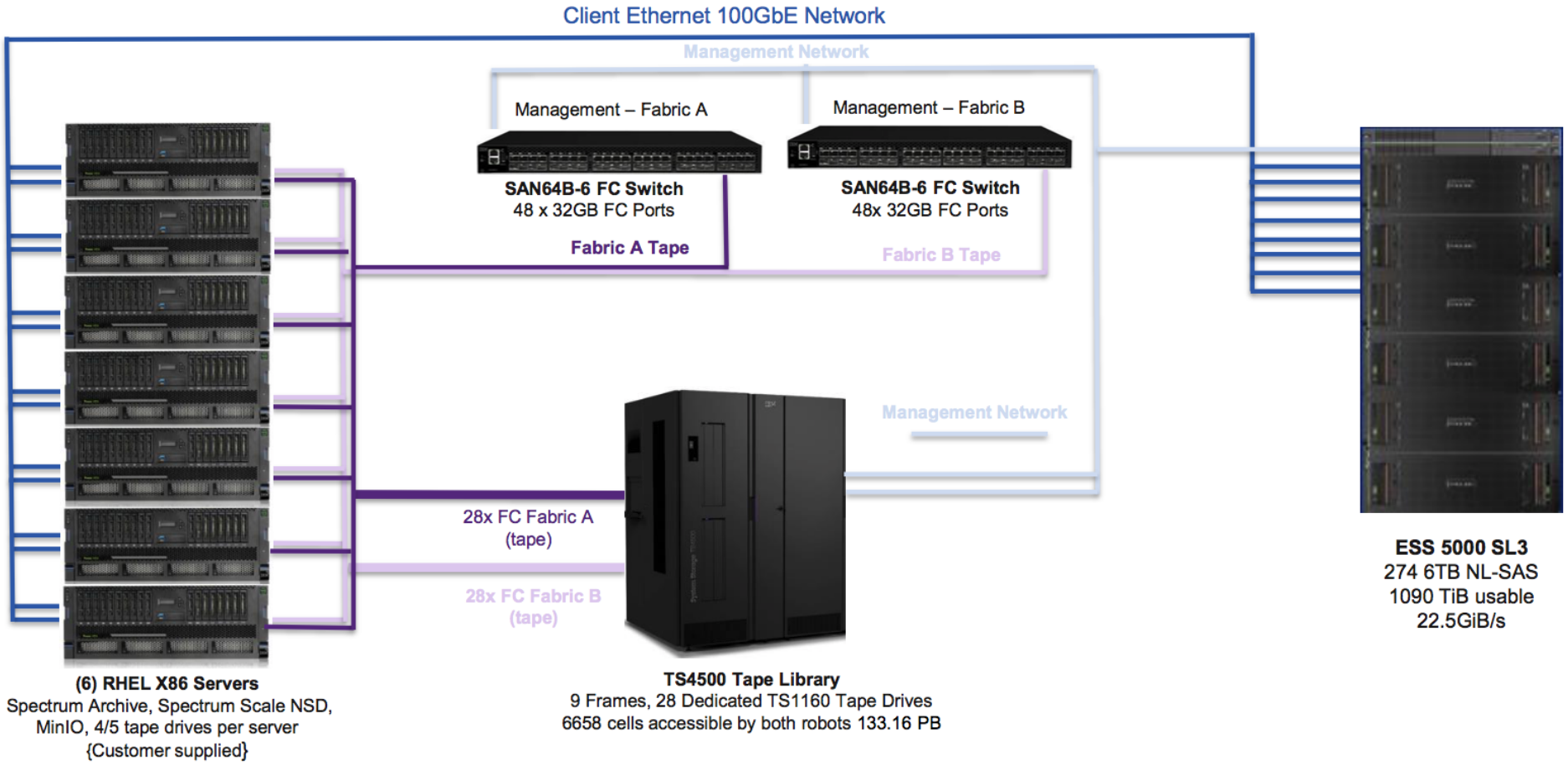
- L'idée est d'utiliser le système comme un HSM (mon image) :
  - Il y a une grande volumétrie sur bande
  - Un buffer disque, les fichiers peuvent être en permanence sur disque (pinned) ou bien temporairement (évincé selon leur popularité)
  - Vu comme un système disque de l'extérieur (rucio, ATLAS DDM)
    - Idéalement serait une QoS intermédiaire entre disque et bande du point de vue de la latence
  - Implémenté avec dCache QoS
- Retour d'expérience :
  - Comme pour le moment rucio / DDM / le job pilot ne sait pas traiter cette QoS intermédiaire, il est difficile de gérer les latences très variables → timeout dans les jobs
  - Les efforts se concentrent maintenant sur l'implémentation de la QoS dans rucio

Résumé

- Utilisation non standard de bandes
  - NESE data center / NET2 :
    - librairie en service, bientôt pour ATLAS mais utilisation « habituelle » d'archivage
  - MÁS à BNL :
    - Sytème « transparent de buffer disque devant une librairie »
    - En l'état actuel difficile à utiliser pour ATLAS, il faut implémenter une QoS dans rucio / DDM
- Difficile de tirer des leçons pour la France même si un site avait accès à une librairie de bande partagée

Backup

Spectrum Archive Architecture ESS



Maximum Capacity: 157 PB
9 Frames, expandable to 18
28 TS1160 Tape Drives, max 11.2 GB/s
ESS-5000: 1.09 PB useable
FC and 100Gb networking

IBM GPFS POSIX interface
IBM Spectrum Archive Library Software
Xrootd with staging
Globus 5 with staging
S3 via MinIO